Exploring Telecom Customer Churn Prediction with Machine Learning

by Md Ishtiaq Ahmed, Israel Gonzalez

Problem Statement

Predicting Customer Churn is an important part of customer relationship management and can help companies improve their customer retention, reduce costs, and improve their financial performance. However, predicting Customer Churn is always a difficult task to do specially for Telecom service providers. Developing a predictive model that accurately identifies customers who are likely to churn from telecom services can help a Telecom company to retain their customers and improve their profit.

Motivations

On the one hand, Ishtiaq worked in a Telecom company for around 8 years. During his tenure, he always saw how they were struggling to retain their customers. It is also difficult to find out a solid reason behind the churn. If they can predict accurately which customers are about churn, then they can take predictive measures to avoid that. On the other hand, Israel had one-year academic experience in Chile in 2013, when he was part of a Business Intelligence Diploma v3 at University of Chile, where Customer Churn was a core business case studied as a very practical type of need where BI and ML solutions can contribute. So, this problem has real-world implications for business performance and profitability which worked as a motivation for us. Also, predicting customer churn is a well-studied problem in machine learning, and there are many techniques and algorithms that can be applied to this problem. By working on this problem, we can apply different techniques learned from the course.

Related works

Predicting Customer Churns in Telecom industry has always been difficult due to the complex behavior of customers and their changing preferences. There were lots of research work has been done, and many are on-going in this area.

One of the research projects we found with the title *“Behavior-Based Telecommunication Churn Prediction with Neural Network Approach”* *(Zhang Y. et al, 2011)* used neural networks to predict customer churn in a Telecom company. Customer service usage information are utilized as the features. Customer churn was predicted using clustering algorithms.

Another research project was *“Intelligent Decision Forest Models for Customer Churn Prediction”* *(Usman-Hamza et al, 2022)*. In this paper, several techniques were used to predict churn including Random Forest algorithm, Functional Tree algorithm, and Logistic Model tree (LMT) algorithm. Conclusions of this study show that this mentioned algorithm gives better results than the classification algorithms like Naïve Bayes (NB) and KNN.

Proposed Solution

1. *Data Analysis*
    a. *Exploratory Data.* This is important to understand the characteristic of data. It will help to identify potential issues with the data and can provide insights about the structure of the data. Which is important to select appropriate machine learning models.
    b. *Cleaning the Dataset.* We proceed to drop unnecessary attributes and samples so that we can manage efficiently the resources to model our problem.
2. *Possible Machine Learning Taks*
    a. *Classification.* It is to predict which customers are going to be churned. By applying classification techniques, we will try to classify the customers in two classes that is churn or not churn.
    b. *Clustering.* We divide the customers in different groups based on their behavior. We will try to analyze if there is any behavioral pattern that can segregate a group of customers.

As part of our *initial solution*, we consider the principal features and run classification algorithms (Perceptron and Logistic Regression) so that we can see the general accuracy to predict customer churn.

Data set

The churn of a telecom customer can depend on many things, and it is challenging to predict customer churn. That is why we were looking for a telecom customer dataset with a large number of instances along with a good amount of features. Finally, we found a dataset from
https://data.world/kishoresjv/telecomchurn/workspace/file?filename=telecom_churn_data.csv

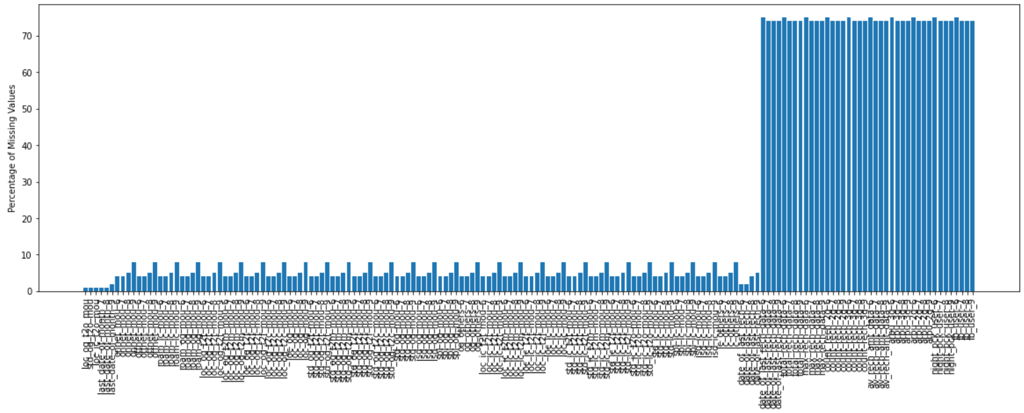
This dataset contains 99,999 instances and 225 features.

Below is a brief description of the dataset:

- Each row of this dataset represents one unique customer
- All the features are telecom customer attributes related to what services they are using, spending on different services, talk time, data usage, recharge amount, data of last usage, date of recharge, types of data pack and many others
- Every attribute data is for four month
- One additional column is added to indicate churn. Customer who did not generate any revenue in the month of September fall under churned customer
- A detail data dictionary is provided in a separate excel sheet.

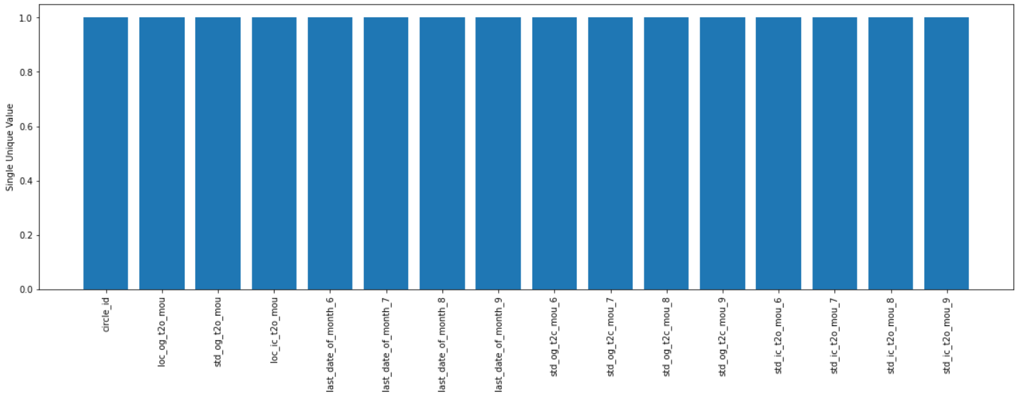
Null value Analysis

From the below figure, we found that there are 40 features where more than 70% value is null. These columns are deleted since they will not add significant value.



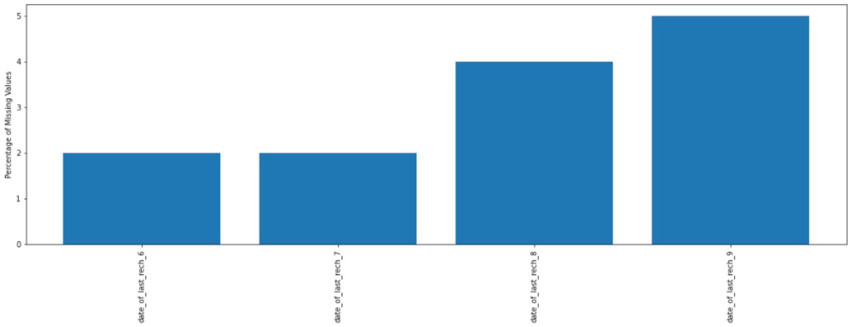
Unique Value Analysis

There are a total of 16 features where only a single unique value is present. These features are deleted from the dataset.



Missing Values Treatment

There are many standard ways to fill up the null values. We have used mean value of each numeric column to fill up null values. After performing the action below is the current status of percentage of null value > 0:



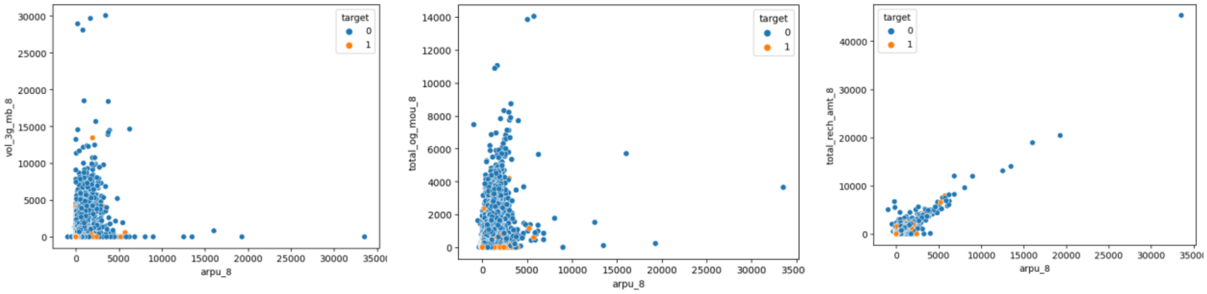
So, there are still some null values present in the date column. Since date values cannot be replaced with mean, we will replace them with zero.

Definition of churn, unbalanced data and nonlinear problem

We have taken for the definition of customer churn that a churn has left the service if in the 9th month he has not consumed calls nor data having consuming data in the previous months.

Following the definition, and mainly for further purposes, we show in our project that the data is unbalanced. This is, 89.91% of the customer stays with service and 10.19% represents the churn regarding our dataset.

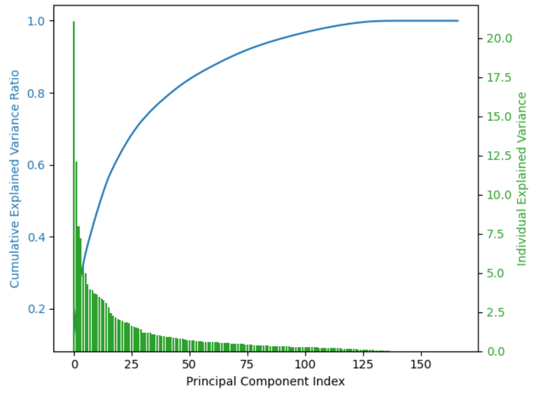
Also, by preparing some scatterplots where we cross arpu_8, vol_3g_mb_8, total_og_mou_8, and total_rech_amt_8; we show that the underlying problem this project deals with is a nonlinear problem because there is not possible to easily with a line divide the class instances.



This conclusion could be important in later stages so that we can refine, for example, the strategy to make a more elaborated technique for extraction attributes such as KPCA.

PCA (Principal Components Analysis) and reducing features

Considering the numerical attributes, a ratio of splitting training/testing of 70:30, and properly standardizing the data, we calculate the covariance matrix, eigenvectors, and eigenvalues. Then, we proceed to take the PCA obtaining the cumulative explained variance and by principal component index, as the following chart demonstrates:



Plot cumulative and individual explained variance vs. principal component index

Seeing this chart, we decide that approximately in 85% of the cumulative explained variance ratio we can just consider 50 out of 167 attributes since they contribute better to the correct solution of their mutual variance.

Initial Solution: Models and general accuracy

To show that we have produced a correct analysis of the data, data selection, and a consistent data extraction, then we apply two classification models, Perceptron and Logistic Regression to our subset using these principal 50 components. The result obtained for the general accuracy we have obtained has improved in comparison is the following:

- Perceptron accuracy (no pca50) = 0.8981
- Perceptron accuracy (with pca50) = 0.9994
- Logistic Regression accuracy (no pca50) = 0.8981
- Logistic Regression accuracy (with pca50) = 0.9995

It is important to note here that this 99.94% of general accuracy, even though is very good, is anomalous and in later stages of our project we will apply techniques to solve this possible overfitting issue like, as we mentioned before, balancing the data.